130-4/22.

COMPONENT PART NOTICE

THIS PAPER IS A COMPONENT PART OF THE FOLLOWING COMPILATION REPORT:

TITLE: Computing Science and Statistics: Proceedings of the Symposium on Interface Critical Applications of Scientific Computing (23rd): Biology, Engineering, Medicine, Speech Held in Seattle, Washington on 21-24 April 1991.

TO ORDER THE COMPLETE COMPILATION REPORT, USE AD-A252 938.

— THE COMPONENT PART IS PROVIDED HERE TO ALLOW USERS ACCESS TO INDIVIDUALLY AUTHORED SECTIONS OF PROCEEDING, ANNALS, SYMPOSIA, ETC. HOWEVER, THE COMPONENT SHOULD BE CONSIDERED WITHIN THE CONTEXT OF THE OVERALL COMPILATION REPORT AND NOT AS A STAND-ALONE TECHNICAL REPORT.

THE FOLLOWING COMPONENT PART NUMBERS COMPRISE THE COMPILATION REPORT:

AD#: AD-P007 096 thru AD#: AD-P007 225

AD#: AD#:

AD#: AD#:

Accesion For	
NTIS CRA&I	☒
DTIC T/B	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	

DTIC ELECTE JUL 23 1992 S A D

This document has been approved for public release and sale; its distribution is unlimited.

DTIC FORM 463 MAR 85

OPI: DTIC-TID

Drug Design : Examining Large Experimental Designs

S. Stanley Young

Glaxo Inc.
Research Triangle Park, NC 27709

Abstract

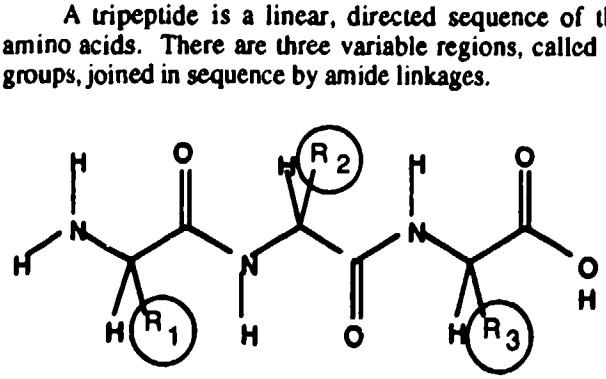
In the course of designing a new drug, thousands of candidate structures could be made and examined by empirical testing. Medicinal chemists would prefer some way of selecting a diverse subset from a list of candidates. Our statistical approach is to use experimental design technology for the selection process and to use computer visualization techniques for examination of the resulting design. A small peptide case is used as an example. The emphasis of this paper is on the value of visualization techniques in understanding the design and in explicating the design to Medicinal Chemists.

Introduction

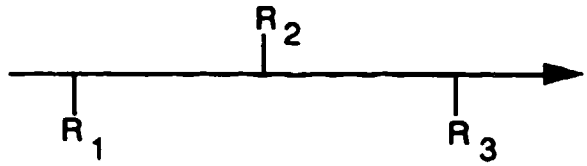
There are countless numbers of molecules that could be made for testing as potential drugs. Ten million different molecules have been made and registered; for most of these molecules that have the characteristics of typical drugs, there are millions of possible modifications. Since it is impossible to make all these molecules, there is a need to create diverse sets of molecules that span the range of possible structures. Hopefully, the "gaps" between the compounds in the design set will be small enough that important compounds are not missed. Our idea is to describe molecules numerically, use statistical experimental design software to create a design set, and examine the resulting design using 3D rotating scattergraph techniques. The process is illustrated using tripeptides.

What is a Tripeptide?

A tripeptide is a linear, directed sequence of three amino acids. There are three variable regions, called side groups, joined in sequence by amide linkages.



There is a beginning amide group, $-NH_2$, and a terminal carboxyl group, COOH. There are three variable regions denoted by R_1, R_2, and R_3 and there is a direction to the molecule. The following diagram captures these features.

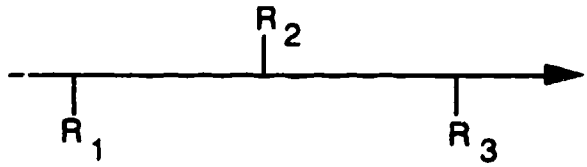


There are 20 naturally ocuring amino acids, so there are 20x20x20=8,000 possible tripeptides. The cost of making enough compound for testing is about $500, so it would cost about four million dollars to make all possible tripeptides. Because this cost is too high and the process would take a long time to complete, it was decided to make a small, diverse set of tripeptides in the hope that a more cost effective discovery process would result.

Numerically Characterize a Tripeptide

Each of the variable regions of a peptide can be described using three numbers. The size can be measured as volume or surface area. Electronic properties can be measured. Also the lipophilicity of the side group can be measured. Lipophilicity is the propensity to dissolve in a water or oil environment. The blood is a water environment, as is the interior of a cell. Between the two is an oily cell membrane. Drugs typically have to pass from blood to the interior of cells so the water/oil relative solubility is important.

To numerically describe a tripeptide we combined these three numerical measures of side group properties across the three positions using linear scales.



Mean	1	1	1	Total
Linear	-1	0	+1	Gradient
Quad	-1	2	-1	Width

Note the three positions from left to right. For each of the three numerical descriptors, size, electronics, and lipophilicity, we created three scores, mean, linear, and quadratic. These scores have physical interpretations. For example, if one adds up the size of each side group at each of the three positions, then the score reflects the total size of the tripeptide. As the tripeptide is directed, the linear component measures a gradient along the tripeptide. Because the R_2 group is typically on the opposite side of the tripeptide from the R_1 and R_3 groups, the size quadratic score measures the width of the tripeptide.

There are three measures of properties of side groups and there are three scores determined for each so there are nine numerical measures of tripeptide properties. In addition to these scores, we computed various interactions among the nine scores to give a total of 34 descriptive variables, ie each of the 8,000 tripeptides was characterized with a vector of 34 numerical descriptors. The problem was to select about 100 tripeptides from the 8,000 so that the resulting set was as diverse as possible.

Experimental Design

There are about 10^{232} ways to select 100 objects from 8,000. We chose to use statistical experimental design software to make this selection. Our problem was much bigger than problems typically attempted using statistical experimental design software, so we had to improvise using various commercially available and internally developed software.

Experimental Design Software

1.	EChip	PC
2.	ACED	VAX or IBM
3.	OPTEX	IBM3090
4.	Inhouse Fortran	IBM3090

Because EChip on the PC would handle only relatively small problems, various iterative strategies were used. For example, one can select a trial design from a small random set of points, say 100 out of 800, do this several times, then make a final selection from the "winners" of each of the trial designs. Solutions on the PC took days to compute. ACED code was obtained from Dr. W. Welch of the University of Waterloo and modified to handle our large problems. We increased memory allocations and in certain instances compiled for a vector processor. We were able to obtain solutions in hours on our mainframes. Vector processing greatly speeded up the selection process. After much effort we were able to obtain a good 82 point design. This design had 55 percent G-optimality. Several designs consisting of 82 randomly selected points were checked. These random designs typically had G-optimality of 1 to 2 percent.

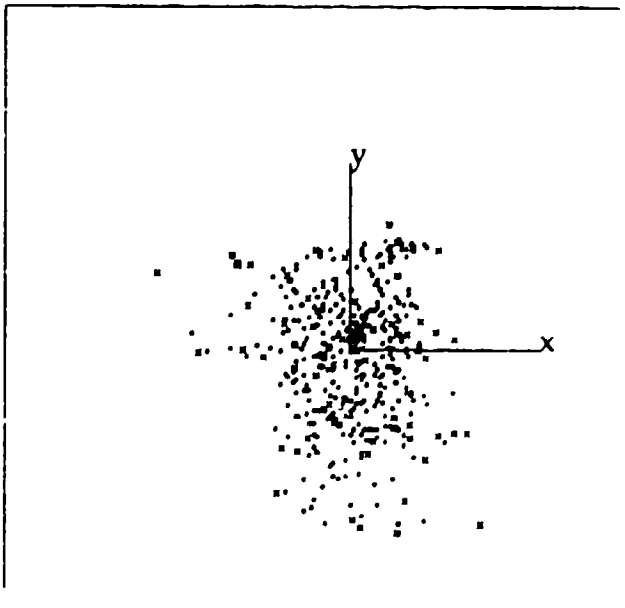
Several comments are in order. Orthogonal polynomials "fold" a dimension. For example, a tripeptide that has a large, small, large R-group in the three positions will be intermediate in size score for the mean polynomial and hence not selected as a vertex, but it will be large for the quadratic polynomial and will be selected as a vertex. The quadratic polynomial folds the size space moving a center point to an extreme point. D-optimal design software selects points that are vertices in a space. An obvious strategy is to select extreme points in the various dimensions as starting points for a design. We are attempting to saturate a low dimension space and do it by creating a higher dimension space that has the right vertices for the lower dimension space.

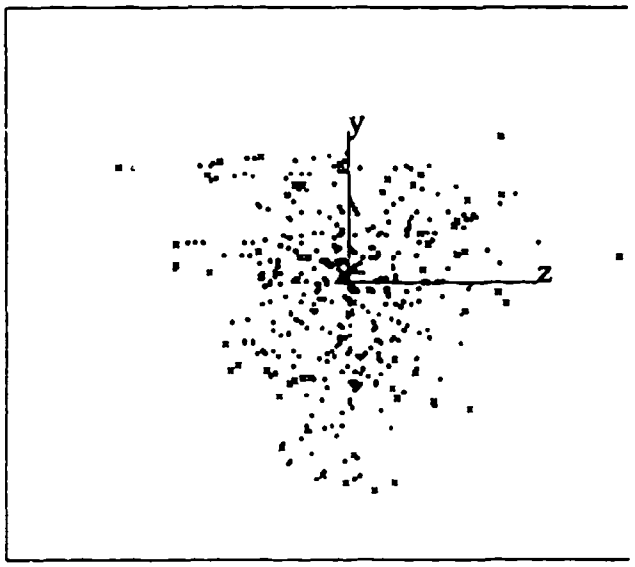
Software for Visualization

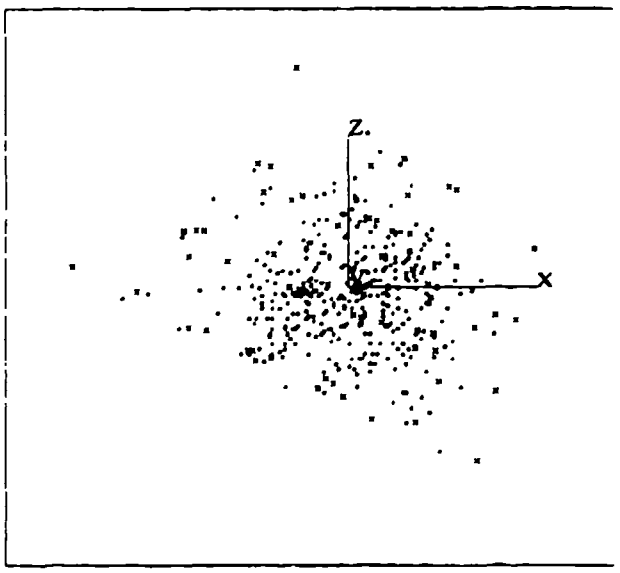
The experimental design software produces an analytical solution to the selection of representative tripeptides. Our resulting design had 82 points in a 34D space. To evaluate this design we used various 3D rotating scattergraph programs. This work was done on a Macintosh and we used MacSpin, Data Desk, and JMP. All three software packages were effective, although each had different features that helped in the visual evaluation of the design.

Our evaluation proceeded as follows. First we selected a random set of 800 points from the 8,000. This was necessary as rotation speed was a function of data set size Next, we added the 82 design points to the data set and marked them with color and/or a distinctive symbol. We then proceeded to look at various 3D projections of the random and design points.

The following figures shows three 90 degree views of the first three dimensions of the data.







In MacSpin we could slice through the cloud of points to examine the number and spacing of design points in planes of the data cloud.

Note that there are about 6,000 ways to select three dimensions from the 34. Also note that if a certain projection looks bad, design points are absent or poorly spaced, then there is no easy way to fix the design. Dropping one design point because it is visually close to another in a certain subspace and adding a point to fill in a void are likely to upset the design in other dimensions. The visualization is reassuring, but it does not offer an easy way to fix a perceived deficient design.

Discussion

Visualization helps assure the statistician that analytical techniques have been correctly employed. With many analytical techniques it can be difficult to detect if gross mistakes are made. It was quite assuring to the statisticians that the points of the final design seemed to saturate the 34D space. To make the 82 tripeptides cost about 50k dollars and took considerable time. Chemists and managers had to evaluate to reasonableness of the effort. Visualization was very effective in showing non-statisticians what was being proposed and some of the limitations, eg the gaps between design points, of the procedure. The collaborators in this project were chemists and Medicinal Chemists tend to think in highly visual ways. 3D rotating scattergraphs were very appealing to them.

Most of this work was done some time ago. In the meantime desktop computers have become much more powerful. Experimental design work could now be done on workstations, particularly if overnight or weekends were available.

The visualization of multiple dimensions is still a problem. With 3D rotation, color and symbols it is possible to get some feel for 4-5D, but we were working in 34D and we wanted to have good assurance of the saturation of 9D in our 34D space. After time consuming visual examination, we became comfortable that we had done a reasonable job, but it did take time and if we had found deficiencies, we would have had no recourse but to start all over again.

Computer Programs

ACED is a copyrighted program of Dr. W.J. Welch.
DataDesk is a trademark of Data Description, Inc.
EChip is a trademark of Expert in a Chip, Inc.
JMP is a trademark of SAS Institute Inc.
MacSpin is a trademark of D^2 Software, Inc.
Optex is a trademark of SAS Institute Inc.

Three graduate students from North Carolina State University worked on this project, Kim Carswell, Kris Latour, and Dan McCaffrey. Their work is gratefully acknowledged. Three professional statisticians also provided insights and software, Randy Tobias, and John Sall of SAS Institute Inc., and William J. Welch of U. of Waterloo.